

Bibliothèques numériques I

Gestion de projet (conception, partage, archivage)

# Modéliser les données

Simon Gabay

Genève



# Remarques introductives

## *In principio erat verbum*

- Importance de la linguistique computationnelle dans les humanités numériques (compter les mots)
- La TEI fondée par l'*Association for Computers and the Humanities*, l'*Association for Computational Linguistics* et l'*Association for Literary and Linguistic Computing*
- D'où l'importance de l'étude des textes avant les autres choses (objets, musique, films...)

## Description et échange des données

- Importance des institutions patrimoniales (musées, bibliothèques, archives)
- Les systèmes d'échange synthétisent les données essentielles à la description
- XML est le moyen privilégié de l'échange de données – et il est lisible par l'être humain
- Autant de raison de s'attarder sur ces formats, entre autres pour des raisons pédagogiques
- Il existe évidemment des systèmes bien plus complexes...

**Dublin core**

## Dublin core

*Dublin Core Metadata Initiative* (DCMI), créée en 1995, Dublin (Ohio, pas Irlande). Elle permet de décrire des documents de manière simple et standardisée

En deux parties:

- *Dublin Core element set*: quinze propriétés
- *Dublin Core metadata terms*: d'autres propriétés supplémentaires

Vous trouverez plus de documentation en ligne:

<https://www.dublincore.org/specifications/dublin-core/profile-guidelines/>

## ***Dublin Core element set***

<b>Nom</b>	<b>Description</b>
Title	Nom donné à la ressource
Creator	Nom de la personne responsable de la création de la ressource
Subject	Thème du contenu
Description	Présentation du contenu
Date	Date de création
Language	Langue du contenu intellectuel
Relation	Référence à une ressource apparentée
Coverage	Couverture spatio-temporelle
Rights	Informations sur les droits associés

## DCMI element set: suites

Nom	Description
Publisher	Organisme de diffusion
Contributor	Personne responsable de contributions au contenu
Type	Nature ou genre
Format	Manifestation physique ou numérique
Identifier	Référence univoque dans un contexte donné (URI, ISBN)
Source	Référence dont la ressource décrite est dérivée (URI)
...	...



## Metadata Terms (extension de l'*element set*)

- dateCopyrighted
- rightsHolder
- created
- issued
- provenance
- isPartOf
- isVersionOf
- hasVersion
- tableOfContents

## Entre vocabulaire et langage

- Dublin core est un vocabulaire du web sémantique
- Il utilisé pour exprimer les données dans un modèle RDF (*Ressource description framework*)
- Il peut être exprimé avec une syntaxe XML ( `.xml` )
- Il peut être exprimé avec une syntaxe Turtle ( `.ttl` )
- Il peut être exprimé avec une syntaxe N-Triples ( `.nt` )

## Plus loin que DC

- *MAchine-Readable Cataloging* (MARC)
- *Metadata Object Description Schema* (MODS, entre DC et MARC)
- *Metadata Encoding and Transmission Standard* (METS, intègre tous les formats: MARC, DC, EAD...)

### Echanger les données

- Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)  
-> [Exemple d'e-codices](#)
- SRU=Search/Retrieve via URL  
-> [Exemple de la BnF](#)

# La TEI: définitions et dérivations

## <MsDesc>

<MsDesc> permet de décrire le manuscrit

- <msIdentifier> pour la cote
- <author> pour l'auteur
- <docDate> pour la date
- <support> pour la description du matériaux (parchemin, vélin...)
- <extent> pour le format (taille, longueur...)
- <condition> pour son état de conservation
- La description peut être extrêmement complexe (mains, enluminures, sceaux, filigranes)
- Description de manuscrit: *Antiphonarium Lausannense. De Sanctis, pars hiemalis. Officium B.M.V. Commune Sanctorum*: sur [www.e-codices.ch](http://www.e-codices.ch)

# La documentation

Il existe des dictionnaires définissant chacun des termes. Cf. <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html> pour la TEI

<msDesc>		Home C Elements
<msDesc> (manuscript description) contains a description of a single identifiable manuscript or other text-bearing object such as an early printed book. <a href="#">[10.1 Overview]</a>		
Module	msdescription — Manuscript Description	
Attributes	<a href="#">att.global</a> (@xml:id, @n, @xml:lang, @xml:base, @xml:space) ( <a href="#">att.global.rendition</a> (@rend, @style, @rendition)) ( <a href="#">att.global.linking</a> (@corresp, @synch, @sameAs, @copyOf, @next, @prev, @exclude, @select)) ( <a href="#">att.global.analytic</a> (@ana)) ( <a href="#">att.global.facs</a> (@facs)) ( <a href="#">att.global.change</a> (@change)) ( <a href="#">att.global.responsibility</a> (@cert, @resp)) ( <a href="#">att.global.source</a> (@source)) <a href="#">att.sortable</a> (@sortKey) <a href="#">att.typed</a> (@type, @subtype) <a href="#">att.declaring</a> (@decls) <a href="#">att.docStatus</a> (@status)	
Member of	<a href="#">model.biblLike</a>	
Contained by	<b>core:</b> <a href="#">add</a> <a href="#">cit</a> <a href="#">corr</a> <a href="#">del</a> <a href="#">desc</a> <a href="#">emph</a> <a href="#">head</a> <a href="#">hi</a> <a href="#">item</a> <a href="#">listBibl</a> <a href="#">meeting</a> <a href="#">note</a> <a href="#">orig</a> <a href="#">p</a> <a href="#">q</a> <a href="#">quote</a> <a href="#">ref</a> <a href="#">reg</a> <a href="#">relatedItem</a> <a href="#">said</a> <a href="#">sic</a> <a href="#">stage</a> <a href="#">textLang</a> <a href="#">title</a> <a href="#">unclear</a> <b>dictionaries:</b> <a href="#">case</a> <a href="#">colloc</a> <a href="#">def</a> <a href="#">dictScrap</a> <a href="#">entry</a> <a href="#">entryFree</a> <a href="#">etym</a> <a href="#">form</a> <a href="#">gen</a> <a href="#">gram</a> <a href="#">gramGrp</a> <a href="#">hom</a> <a href="#">hyph</a> <a href="#">iType</a> <a href="#">lang</a> <a href="#">lbl</a> <a href="#">mood</a> <a href="#">number</a> <a href="#">orth</a> <a href="#">per</a> <a href="#">pos</a> <a href="#">pron</a> <a href="#">re</a> <a href="#">sense</a> <a href="#">stress</a> <a href="#">subc</a> <a href="#">syll</a> <a href="#">tns</a> <a href="#">usg</a> <a href="#">xr</a> <b>drama:</b> <a href="#">camera</a> <a href="#">caption</a> <a href="#">castList</a> <a href="#">epilogue</a> <a href="#">performance</a> <a href="#">prologue</a> <a href="#">set</a> <a href="#">sound</a> <a href="#">tech</a> <a href="#">view</a> <b>figures:</b> <a href="#">cell</a> <a href="#">figDesc</a> <a href="#">figure</a> <b>header:</b> <a href="#">change</a> <a href="#">handNote</a> <a href="#">licence</a> <a href="#">rendition</a> <a href="#">scriptNote</a> <a href="#">sourceDesc</a> <a href="#">tagUsage</a> <a href="#">taxonomy</a> <b>iso-fs:</b> <a href="#">fDescr</a> <a href="#">fsDescr</a> <b>linking:</b> <a href="#">ab</a> <a href="#">seg</a> <a href="#">standOff</a> <b>msdescription:</b> <a href="#">accMat</a> <a href="#">acquisition</a> <a href="#">additions</a> <a href="#">collation</a> <a href="#">condition</a> <a href="#">custEvent</a> <a href="#">decoNote</a> <a href="#">filiation</a> <a href="#">foliation</a> <a href="#">layout</a> <a href="#">msItem</a> <a href="#">musicNotation</a> <a href="#">origin</a> <a href="#">provenance</a> <a href="#">signatures</a> <a href="#">source</a> <a href="#">summary</a> <a href="#">support</a> <a href="#">surrogates</a> <a href="#">typeNote</a> <b>namesdates:</b> <a href="#">climate</a> <a href="#">event</a> <a href="#">location</a> <a href="#">object</a> <a href="#">occupation</a> <a href="#">org</a> <a href="#">person</a> <a href="#">personGrp</a> <a href="#">persona</a> <a href="#">place</a> <a href="#">population</a> <a href="#">state</a> <a href="#">terrain</a> <a href="#">trait</a> <b>spoken:</b> <a href="#">broadcast</a> <a href="#">scriptStmt</a> <a href="#">u</a> <a href="#">writing</a> <b>textcrit:</b> <a href="#">lem</a> <a href="#">rdg</a> <a href="#">witness</a> <b>textstructure:</b> <a href="#">argument</a> <a href="#">body</a> <a href="#">div</a> <a href="#">div1</a> <a href="#">div2</a> <a href="#">div3</a> <a href="#">div4</a> <a href="#">div5</a> <a href="#">div6</a> <a href="#">div7</a> <a href="#">docEdition</a> <a href="#">epigraph</a> <a href="#">imprimatur</a> <a href="#">postscript</a> <a href="#">salute</a> <a href="#">signed</a> <a href="#">titlePart</a> <a href="#">trailer</a> <b>transcr:</b> <a href="#">damage</a> <a href="#">metamark</a> <a href="#">mod</a> <a href="#">restore</a> <a href="#">retrace</a> <a href="#">secl</a> <a href="#">supplied</a> <a href="#">surplus</a> <b>verse:</b> <a href="#">rhyme</a>	
May contain	<b>core:</b> <a href="#">head</a> <a href="#">p</a> <b>linking:</b> <a href="#">ab</a> <b>msdescription:</b> <a href="#">additional</a> <a href="#">history</a> <a href="#">msContents</a> <a href="#">msFrag</a> <a href="#">msIdentifier</a> <a href="#">msPart</a> <a href="#">physDesc</a>	
Note	Although the <a href="#">msDesc</a> has primarily been designed with a view to encoding manuscript descriptions, it may also be used for other objects such as early printed books, fascicles, epigraphs, or any text-bearing objects that require substantial description. If an object is not text-bearing or the reasons for describing the object is not primarily the textual content, the more general <a href="#">object</a> may be more suitable.	

## **<MsDesc> +**

- "Détournement" (ou plus précisément "changement de sémantisme") de `<msDesc>`
- Bibliographie matérielle, pour les catalogues de livres (anciens)
- Pour décrire le support des inscriptions épigraphiques
- Description d'épigraphie (cf. [ISic0298](#))

## Attention à la documentation

Il faut toujours (ou presque) suivre la documentation en anglais

Définition en anglais:

**<msDesc>** (manuscript description) contains a description of a single identifiable manuscript or other text-bearing object such as an early printed book.

Définition en français

**<msDesc>** (description d'un manuscrit) contient la description d'un manuscrit individuel



**LIDO**

# Lightweight Information Describing Objects

- C'est un format d'échange de données
- Il permet de décrire les objets et les ressources numériques (images, textes, sons, vidéos)
- 14 groupes d'informations, dont 3 sont obligatoires
- 5 types de groupes d'information

Plus d'information ici: <https://lido-schema.org/schema/v1.1/lido-v1.1.html>

## LIDO 1: classification

1. **Object/Work type** (classification)
2. **Classification** (style, forme, âge...)

## LIDO 2: événements

3. **Event set** (création, exposition... On y reviendra)

## LIDO 3: relations

4. **Subject set** (objet, bâtiments, personnes dans l'œuvre)
5. **Related Works**

## LIDO 4: identification

6. **Title/Name**
7. **Inscriptions** (transcription et ou description)
8. **Repository/location** (institution et numéro d'inventaire)
9. **State/Edition**
10. **Object description**
11. **Measurements**

## LIDO 5: Administration

12. **Rights**
13. **Record**
14. **Ressources**

## Autres formats pour les musées

- museumdat ([www.museumdat.org](http://www.museumdat.org))
- SPECTRUM XML (<http://www.collectiontrust.org.uk/spectrum>)
- CIDOC-CRM (<http://www.cidoc-crm.org/>)

## **Linked data (CIDOC et autres)**

# LIDO et CIDOC-CRM

Lien avec CIDOC-CRM (*Conceptual reference model*)

- LIDO est un format de description (comme DC)
- CIDOC CRM est un modèle conceptuel (comme FRBR)
- Ce modèle de données est "orienté événement"

Ce modèle "orienté événement" décrit les événements de la vie d'un objet pour décrire ce dernier *via*:

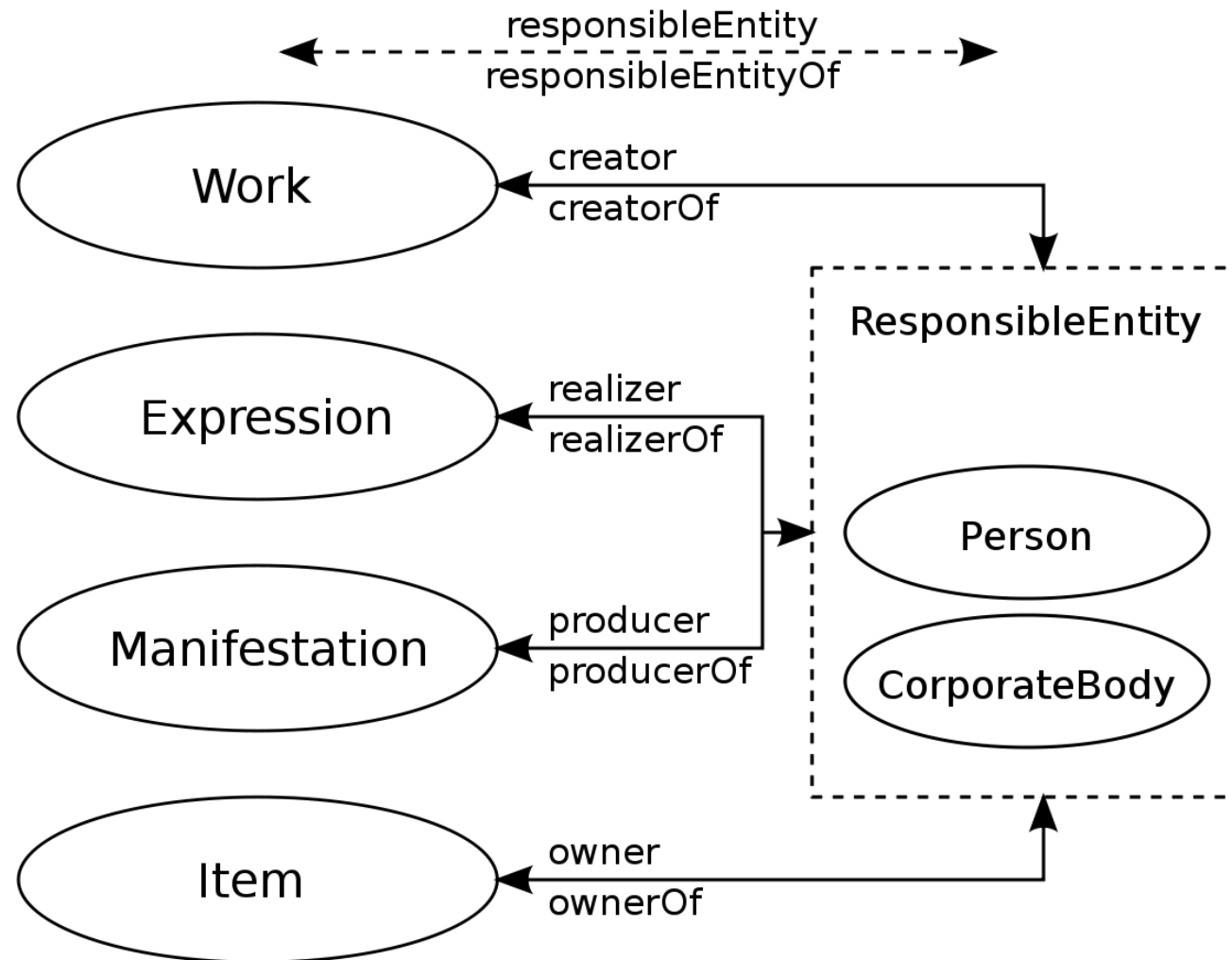
- un/des agent(s)
- une date ou un intervalle dans le temps
- un lieu
- un type d'événement

## FRBR

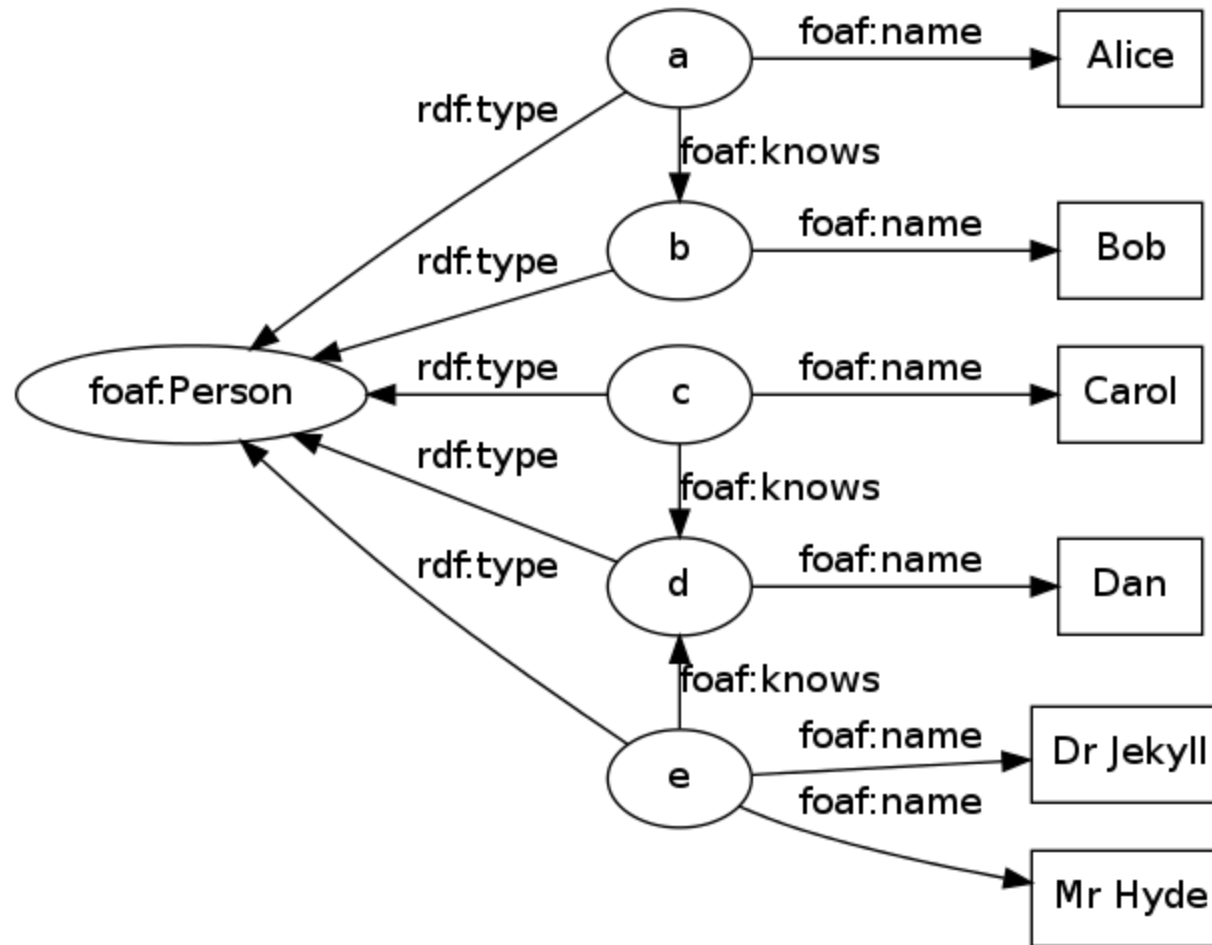
- Œuvre : une création intellectuelle ou artistique déterminée (par exemple : *La Divine Comédie*)
- Expression : une réalisation de cette création intellectuelle (par exemple, l'édition de Petrocchi, la traduction de Risset)
- Manifestation : la matérialisation d'une expression (la traduction de Risset publiée chez GF)
- Item : un exemplaire isolé d'une manifestation (par exemple, la traduction de Risset publiée chez GF publiée à la BGE).



# FRBR



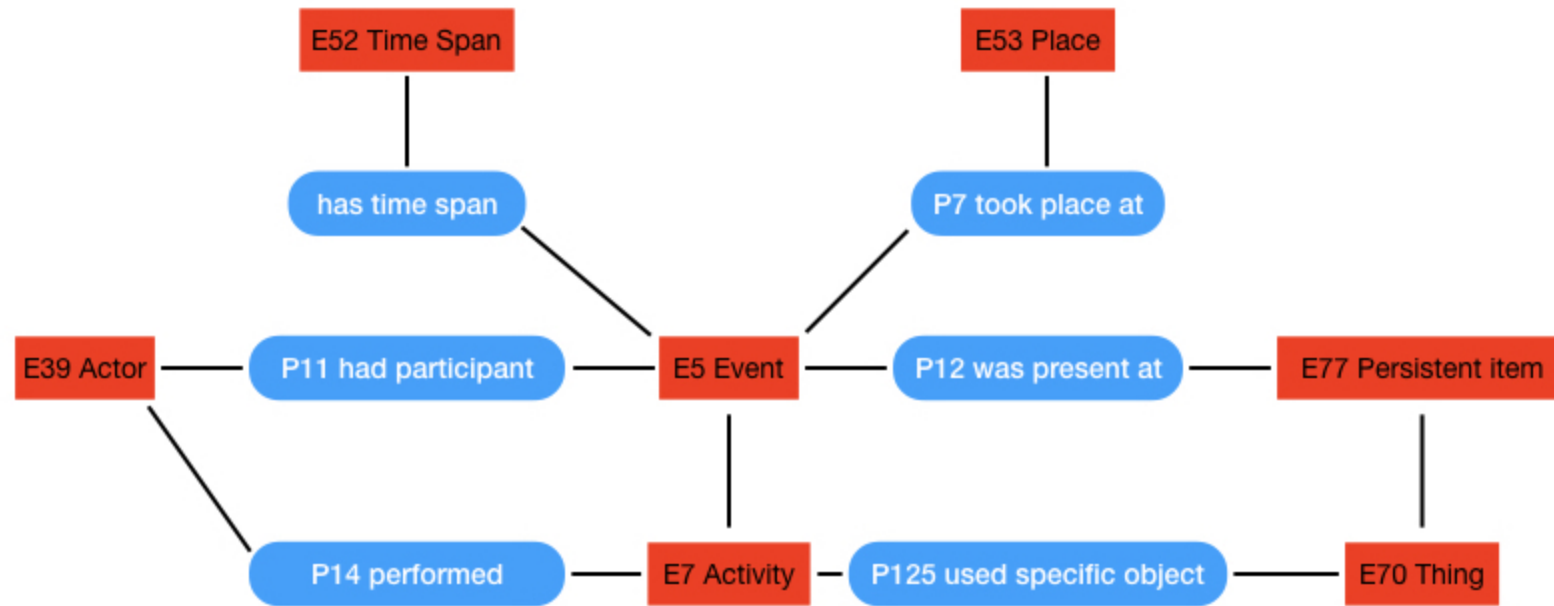
# FOAF



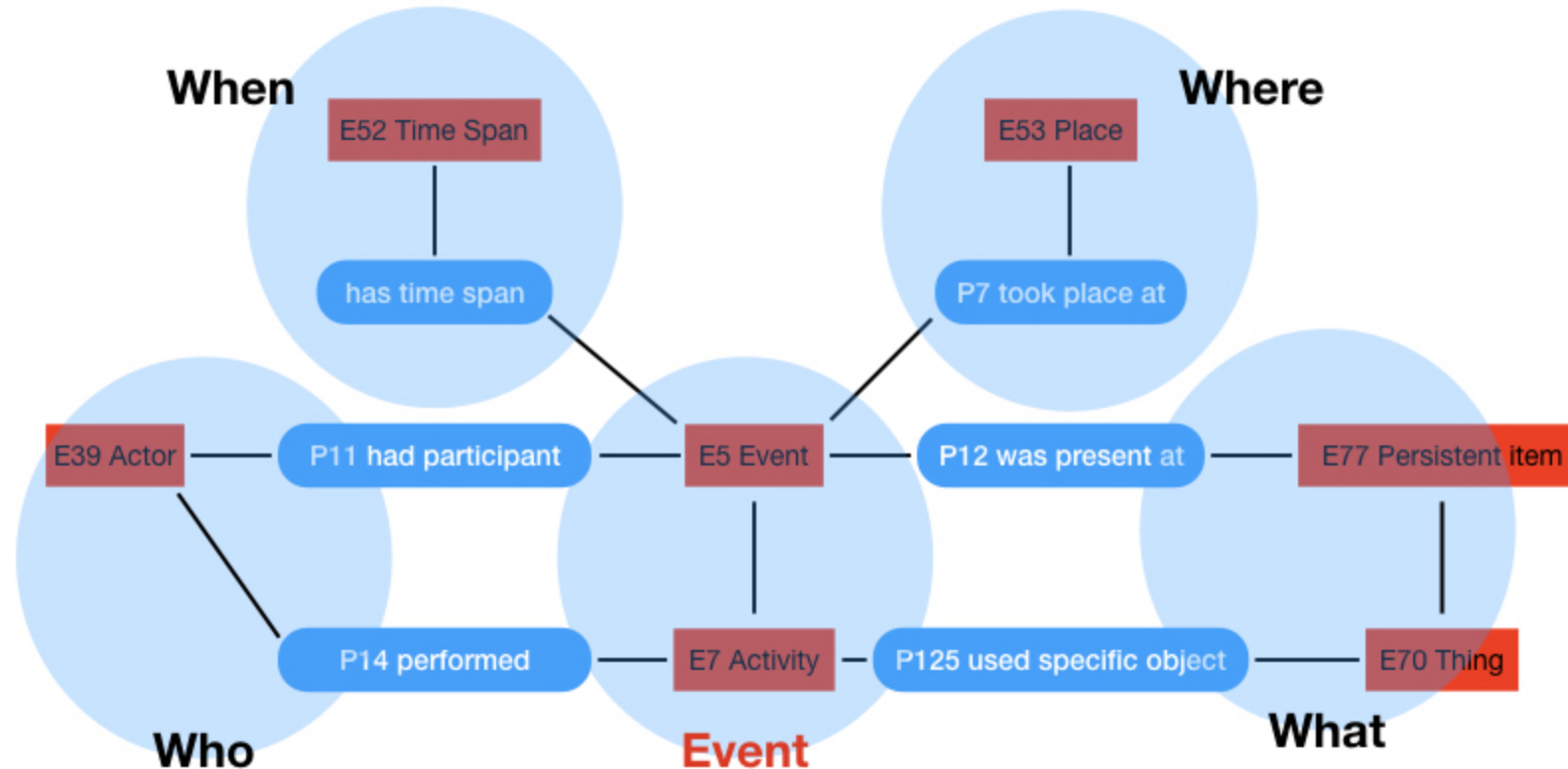
# Un modèle "orienté événement"

- Creation
- Modification
- Part addition
- Part removal
- Excavation
- Acquisition
- Finding
- Exhibition
- Move
- Restoration
- Loss
- Destruction

# CIDOC-CRM simplifié



## Les quatre W (*who, what, when, where*)



# Exemple: Le *Monument à Balzac* de Rodin

